U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
OFFICE OF SCIENCE AND TECHNOLOGY
METEOROLOGICAL DEVELOPMENT LABORATORY

MDL OFFICE NOTE 14-1

ON BIAS CORRECTING MOS WIND SPEED FORECASTS

Bob Glahn, David Rudack, and Bruce Veenhuis

August 2014

ON BIAS CORRECTING MOS WIND SPEED FORECASTS

Bob Glahn, David Rudack, and Bruce Veenhuis

1. INTRODUCTION

MOS temperature, dewpoint, and wind speed forecasts are produced by multiple "linear" regression; the predictand is related to a combination of numerical model, and possibly other, predictors. While the relationship is linear in the specific predictors, these predictors can be non-linear combinations of model and other variables, making the predictand relationship to the original model predictors nonlinear, where the non-linearity has been designed to make meteorological sense. In practice, these forecasts tend to have some bias on both a monthly and seasonal basis. Regression equations that produce MOS forecasts give unbiased estimates over the period of the developmental sample, but they may have bias over intervals within that developmental sample and over other samples including future forecasts. Given there are no changes in the system producing the forecasts (e.g., numerical model, method of data collection, etc.), the bias on a seasonal basis is expected to be small, but the overall synoptic patterns are different from year to year, so biases may occur.

Several bias correction methods have been reported in the literature (e.g., Yussouf and Stensrud 2007; Woodcock and Engel 2005) that largely eliminate forecast bias; however, modification to the forecasts to correct bias can lead to larger mean absolute errors (MAE) and degrade other metrics used to evaluate the quality of the forecasts. Most studies have not dealt with MOS operational forecasts. Glahn (2012; 2014) tested for temperature and dewpoint a method that has been used at the National Centers for Environmental Prediction (NCEP) for several years (Cui et al. 2012) called decaying average and found that it not only improved bias but either improved or did not degrade other performance metrics. This method was compared to a regression method of correction that is employed within the Boise Verify software in use at National Weather Service (NWS) field offices, and it was found that the decaying method was better and easier to implement (Glahn 2013).

While wind speed is a quasi-continuous variable like temperature and dewpoint, there are questions concerning the applicability of the decaying average method to wind speed. The distribution of wind speed is quite different from the quasi-normal distributions of temperature and dewpoint. The Meteorological Development Laboratory (MDL) has found in the past that regression estimates poorly fit the high end of the distribution (i.e., the higher speeds). In fact, an "inflation factor" has been applied to produce forecasts that more nearly match the observed distribution since 1975 (Schwartz and Carter 1982; Jacks, et al. 1990). Even though this inflation actually increases the MAE and mean square error (Gilhousen et al. 1979), other scores deemed more important, such as threat score of high wind, are improved.

"Inflation" was proposed by Isadore Enger and first applied by Klein et al. (1959). Inflated forecasts are obtained by subtracting the developmental sample mean from the regression estimate, dividing the difference by the (multiple) correlation coefficient, and adding the result to the sample mean. MDL found that this worked well for forecasts above the mean, but those below the mean were too weak, so the established practice is to "partially inflate" by using the procedure on only those regression estimates above the mean. Inflation, either full or partial, will increase the variance of the forecasts and increase the mean square error [see Glahn and Allen (1966) for details.]

In this note, we document results of bias correcting with the decaying average algorithm a sample of operational MOS 10-m wind forecasts based on NCEP's Global Forecast System (Caplan et al. 1997) and also experimental MOS forecasts developed on a reforecast dataset (Hamill et al. 2013).

2. DECAYING AVERAGE ALGORITHM

An algorithm for removing bias, called decaying average, has been applied at NCEP to model output since 2006 (Cui et al. 2012). To implement the algorithm, one has only to carry forward a delta d and apply it to the current forecast. Then to prepare for the next forecast cycle, the delta would be updated by:

$$d\,(t+1) = (1-\alpha)\; d\,(t) + \alpha(F - O)(t)$$

where d(t+1) is the delta to apply at time t+1, d(t) is the delta applied at time t, F is the forecast "verified" by the observation O at time t, and α is the weight to apply to the most recently calculated forecast error F - O at time t. There would optimally be a specific delta for each station as well as forecast projection. When F - O is missing, zero can be assumed (see below).

In an operational setting, the modification to the MOS forecast does not have to be made until the observation is available. Therefore, the delta for the next forecast can incorporate the error of the most recent forecast verifying at that time.

3. APPLICATION TO OPERATIONAL MOS WIND SPEED FORECASTS

A choice of a value of α has to be made. NCEP uses α =0.02.[1] Glahn (2012; 2014) tested values of 0.025, 0.050, 0.075, and 0.100 for temperature and dewpoint and found that values of 0.025 and 0.050 were best overall, and the difference in results for those values was not great. While higher values were more effective in removing the bias both long term and short term, the other performance metrics were generally degraded for higher values.

As a practical matter, there are situations where a station will not report for a considerable length of time, or may stop altogether. MOS forecasts continue because they are based on model data that are available. If the delta computed from the time of the last observation were continued, it would likely become inappropriate. To address this potential problem, when the current error could not be computed, it was considered to be zero, so that the decayed average would drift toward zero. In other words, if the past short-term bias is not known, there can be no correction for it. (Biases from surrounding stations could be consulted, but that was not addressed and is likely not worth the effort.)

As another practical matter, in an operational, automated system, the unexpected can happen, and in the situation studied here the MOS forecast or the observation could be highly erroneous, making the computed error disastrous. (There may be other error checks in the system, but they

[1] Cui 2012, personal communication.

are likely not stringent enough to alleviate a problem here.) In order to avert potential disaster, a cap was put on the forecast error that is exactly 20 kt for a 24-h forecast and 40 kt for an 11-day forecast, with the cap defined by a linear line between those two projections and extending on either end as necessary. (For the projections used here, there are no projections outside those limits.) The large differences are still used, but they are capped. That is, if a 55 kt error occurred at 11 days, it would be used as 40 kt. This would still be a shock to the correction algorithm, but not disastrous.

We tested the decaying average method with a value of α equal to 0.04. We chose this value after carefully considering the results for different values for temperature and dewpoint. The metrics used were not much different for values of 0.025 and 0.05, and were overall not as good on either side of these values. We tested on a sample starting January 1, 2011, and ending May 15, 2012. The algorithm was run with a cold start (i.e., a delta of zero) for each station and projection and continued uninterrupted until the end. Then we verified the cool season October 1, 2011, through March 31, 2012. As stated earlier, these forecasts are based on the GFS (Caplan et al. 1997). The stations used were in the contiguous United States, the same 1,319 stations used previously in testing for temperature and dewpoint.

Figure 1 shows that bias correction of the MOS operational forecasts reduced the overall bias over the 2011-2012 cool season. This is the expected result; however, a low bias is not necessarily a characteristic the MOS wind speed forecasts should have. The partial inflation should produce an overall positive bias in the MOS forecasts, and that is indicated in Fig. 1. Because only about half of the forecasts (those above the dependent sample mean) are increased, and those near the mean by not much, the overall bias is small and seems plausible. Because the correlation coefficient decreases with projection, more inflation occurs at longer projections, so it is not surprising that the bias increases with increasing projection.

Along with reducing the bias, the procedure also gave a lower MAE (see Fig. 1) and mean square error (not shown). This is also expected. The regression equations produce the minimum mean square error (MSE) on the developmental sample, and inflation will increase that MSE.

In routine MDL verification, wind speed is categorized for purposes of computing Heidke Skill score and Probability of Detection (PoD). The six categories are shown in Table 1.

Table 1. Definition of wind speed categories.

Category Number	1	2	3	4	5	6
Speed Range (kt)	< 5	≥ 5 and <10	≥ 10 and <15	≥ 15 and <20	≥ 20 and <25	≥25

Figure 2 is surprising. For the 1-, 2-, and 3-day forecasts, the bias of winds ≥ 25 kt is well above unity.[2] However, by day 6, the bias is below unity and as low as 0.5 at day 11. The bias of winds ≥ 20 kt has a different behavior, and is below unity for all projections. The bias corrected forecasts have lower biases than the uncorrected ones as expected, and are considerably below unity. This is not the characteristic we have been striving for. The behavior

[2] Bias for a forecast treated as continuous for verification is defined as the average arithmetic error and has a baseline of zero. When a variable is treated as categorical and counts are involved, bias is defined as the number forecast divided by the number observed, and has a baseline of unity.

for winds ≥ 25 kt shown here for the cool season is the same for the summer season of 2011 (not shown).

It is puzzling why the bias for winds ≥ 25 kt changed so much by projection for inflated forecasts. In looking back at some old wind verifications, it appears the stronger winds were also underforecast. This indicates the inflation procedure, while still very helpful, may not be as effective as had been thought.

Figure 3 indicates the bias was below unity for wind speed forecasts < 5 kt. These forecasts should mostly be below the mean, and therefore not modified by partial inflation. Bias correction increases the bias and brings it closer to unity.

The PoD is the fraction of the time observed strong winds were correctly forecast. It is a measure of accuracy, and the drop with projection shown in Fig. 4 reflects the loss in accuracy. Figure 4 also shows that the bias correction reduced the PoDs of strong winds by a substantial amount at all projections.

Figure 5, similar to Fig. 4, shows that bias correcting the forecasts decreased the Critical Success Index (CSI) (also known as the Threat Score) for speeds ≥ 20 kt and ≥ 25 kt, except for speeds ≥ 25 kt at short projections. Decreasing the accurate forecasting of strong winds is not a desirable characteristic of a postprocessing system.

The bias correction did not change the fraction correct or Heidke Skill Score (SS) by much at short projections (see Fig. 6), but generally improved them at long projections. This is because the bias correction puts more forecasts into the more frequent, easier to forecast, categories. These scores relate to the complete 6 X 6 forecast/observed contingency table. The number correct treats each category the same and does not consider near misses. The Heidke SS is the fraction of possible improvement of the number of correct forecasts of the system being verified over a random forecast, given the observed frequencies. It, too, heavily weights the more frequent categories and gives no credit for near misses. The Gerrity SS (Gerrity 1992) is an equitable score that gives high weight to a rare category and considers the closeness (in number of categories) of the forecast to the observed. While this is not the only such equitable score, and the assumptions underlying the calculation of the utility matrix it uses are somewhat arbitrary, it seems a reasonable attempt to measure the overall goodness of a set of forecasts, taking into account the importance of rare events. Figure 6 shows that bias correcting decreases the Gerrity SS at all projections.

4. APPLICATION TO ENSEMBLE-BASED EXPERIMENTAL WIND SPEED FORECASTS

To further illustrate the effects of the decaying average algorithm on wind speed forecasts, we created and verified MOS forecasts based on the GEFS reforecast ensemble dataset created by Hamill et al. (2013). This dataset is an 11-member ensemble initialized once a day at 0000 UTC. Experimental warm and cool season MOS wind speed equations were developed for projections 48, 120, and 192 hours (8 days) for each of 334 stations in the contiguous United States, Hawaii, Alaska, and Puerto Rico. These 334 stations were carefully selected by MDL for their quality of observations and uniform spatial density and have been used for testing and verification for a number of years.

The 5-year developmental sample for the cool season spanned the months October through March for the years 2007 through 2012; the warm season was comprised of April through September of 2008 to 2012. The development paralleled that for development of the operational equations as to process, selection of predictors, etc. The predictors were the means of the 11 members of the ensemble, rather than from individual members, a technique used previously (e.g., Glahn et al. 2009; Wagner and Glahn 2010; Veenhuis 2013) and preferred by Unger et al. (2009). The test period for each season immediately followed the last season of the developmental period, being October 2012 through March 2013 for the cool season and April through September 2013 for the warm season.

Figures 7, 8, and 9 show for 48-, 120-, and 192-h forecasts, respectively, for the cool season test sample the MAE (across the top) for (1) the (raw) GEFS forecasts, (2) the bias corrected (BC) GEFS forecasts, (3) the (uninflated) MOS forecasts, (4) the partially inflated MOS forecasts, and (5) the bias corrected, partially inflated MOS forecasts. Notably, bias correcting raw model output reduced the MAE, and MOS gave further improvement. Inflating the MOS forecasts increased the MAE as expected, and bias correcting the inflated forecasts decreased the MAE.

The bar graphs in these figures show the sample counts (frequencies) in each of the six categories of wind speed shown in Table 1 (note the ordinate scales are different for the different categories) for each of the five forecast systems enumerated and described above; in addition, the observed frequencies are shown. Of the five categories, those for the two higher categories are the most important. An obvious conclusion is that both the GEFS and uninflated MOS forecasts had far too few strong winds, and the effect was much more pronounced for the longer projections. Consistent with what was found for operational MOS forecasts, described in the previous section, bias correcting these inflated experimental forecasts reduced the frequency of strong wind forecasts. Also consistent with the operational MOS forecasts, inflation does not produce as many strong winds as are observed at the longer projections.

Figures 10, 11, and 12 are the same as the previous three, except for the warm season. The conclusions as to MAE are the same as for the cool season. The underforecasting of strong winds is even more pronounced in the warm season than the cool season, which is reasonable because of synoptic considerations.

5. SUMMARY AND CONCLUSIONS

The decaying average algorithm has been applied to operational MOS wind speed forecasts and to experimental MOS forecasts based on an ensemble reforecast dataset. The operational forecasts have been partially inflated to increase the higher wind speeds. This is necessary for them to be operationally useful. The inflation increases the MAE and overall bias, but improves scores that emphasize the skill and accuracy of strong winds.

The scores computed and shown for one 6-month cool season of MOS forecasts were about as expected, except that the bias dropped off unexpectedly with projection. The reason for this is not known. It is possible the GFS model winds, on which MOS is largely based, had lower biases at longer projections in the test sample than in the developmental sample. However, the same behavior was present with the experimental GEFS-based forecasts.

The bias correction behaved about as expected; the overall bias was generally improved, but scores that emphasized the stronger winds—the most important ones—were worse. It is concluded that this simple bias correction method applied to inflated MOS forecasts would decrease their usefulness to forecasters and other users.

If bias correction were to be done, it should be done before the inflation step, and then inflation applied. This should help with emphasizing stronger winds, and the overall bias might be improved. However, the bias of wind speed forecasts is a rather small component of the total error, especially at the longer projections.

For the experimental MOS equations based on the ensemble means of the GEFS reforecast dataset, we computed MAEs and forecast frequencies in several categories of wind speed. The forecast frequencies, when compared to observed frequencies, indicate category bias. The primary conclusion is that raw model winds, either bias corrected or not, and MOS forecasts based on them are extremely low biased for winds $\geq$ 25 kt and even $\geq$ 20 kt, especially for the longer projections. Such forecasts would likely not be useful to anyone.

In keeping with the results on operational MOS forecasts, the low bias for strong winds is present in the longer range forecasts. This brings into question the long held opinion that bias correction produces as many strong winds as are observed. This belief has theoretical justification, but only for normally distributed variables. Because of the fat right tail of wind speed distributions, it is possible the partial inflation would give better results if done above the median rather than above the mean.

ACKNOWLEDGMENTS

We are indebted to Tamarah Curtis who was very helpful in providing details on graph production.

REFERENCES

Caplan, P., J. Derber, W. Gemmill, S.-Y. Hong, H.-L Pan, and D. Parrish, 1997: Changes to the 1995 NCEP operational medium-range forecast model analysis-forecast system. *Wea. Forecasting*, **12**, 581-594.

Cui, B., Z. Toth, Y. Zhu, and Hou, D., 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396-410 .

Gerrity, J. P., 1992: A note on Gandin and Murphy's Equitable Skill Score. *Mon. Wea. Rev*., **120**, 2709-2712.

Gilhousen, D. B., J. R. Bocchieri, G. M. Carter, J. P. Dallavalle, K. F. Hebenstreit, G. W. Hollenbaugh, J. E. Janowiak, and D. J. Vercelli, 1979: Comparative verification of guidance and local aviation/public weather forecasts—No. 5. *TDL Office Note* **79-2**, Techniques Development Laboratory, National Weather Service, NOAA, U. S. Department of Commerce, 73 pp.

Glahn, B., 2012: Bias correction of MOS temperature and dewpoint forecasts. *MDL Office Note* **12-1**, Meteorological Development Laboratory, National Weather Service, NOAA, U.S. Department of Commerce, 33 pp.

_____, 2013: A comparison of two methods of bias correcting MOS temperature and dewpoint forecasts. *MDL Office Note* **13-1**, Meteorological Development Laboratory, National Weather Service, NOAA, U.S. Department of Commerce, 22 pp.

_____, 2014: Bias correction of MOS temperature and dewpoint forecasts. *Wea. Forecasting* (in press).

_____, and R. A. Allen, 1966: A note concerning the "inflation" of regression forecasts. *J. Appl. Meteor*., **5**, 124-126.

_____, M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, and B. Schuknect, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev*., **137**, 246-268.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bul. Amer. Meteor. Soc*., **94**, 1553-1564.

Jacks, E., J. B. Bower, V. Dagostaro, J. P. Dallavalle, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maximum/minimum temperature, probability of precipitation, cloud amount, and surface wind. *Wea. Forecasting,* **5**, 128-138.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor*., **16**, 672-682.

Schwartz, B. E., and G. M. Carter, 1982: An evaluation of a modified speed enhancement technique for objective surface wind forecasting. TDL *Office Note* ***82-1***, Techniques Development Laboratory, National Weather Service, NOAA, U.S. Department of Commerce, 10 pp.

Unger, D. A., H. Van Den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Mon. Wea. Rev*., **137**, 2365-2379.

Veenhuis, B., 2013: Spread calibration of ensemble MOS forecasts. *Mon. Wea. Rev*., **141**, 2467-2482.

Wagner, J., and B. Glahn, 2010: Ensemble MOS forecasts from multiple models. Preprints *20th Conference on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA, Amer. Meteor. Soc., **7.5**.

Woodcock, F., and C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting,* **20**, 101-111.

Yusssouf, N., and D. J. Stensrud, 2007: Bias-corrected short-range ensemble forecasts of near-surface variables during the 2005/06 cool season. *Wea. Forecasting*, **22**, 1274-1286.

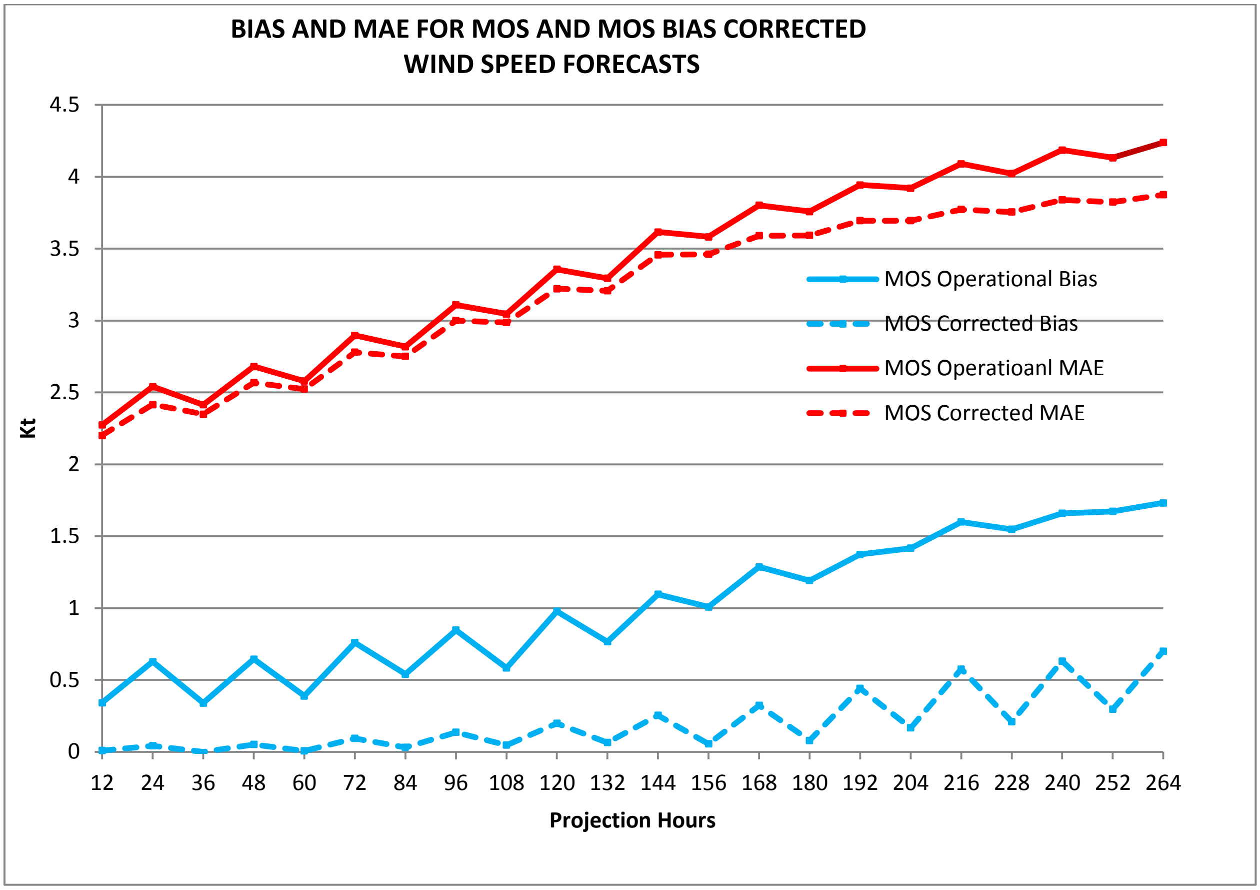

Figure 1. Bias and MAE of MOS operational, partially inflated forecasts and those forecasts bias corrected for the 6-month test period October 1, 2011, through March 31, 2012. Bias is defined as forecast minus observed. Projections are shown at 12-h intervals.

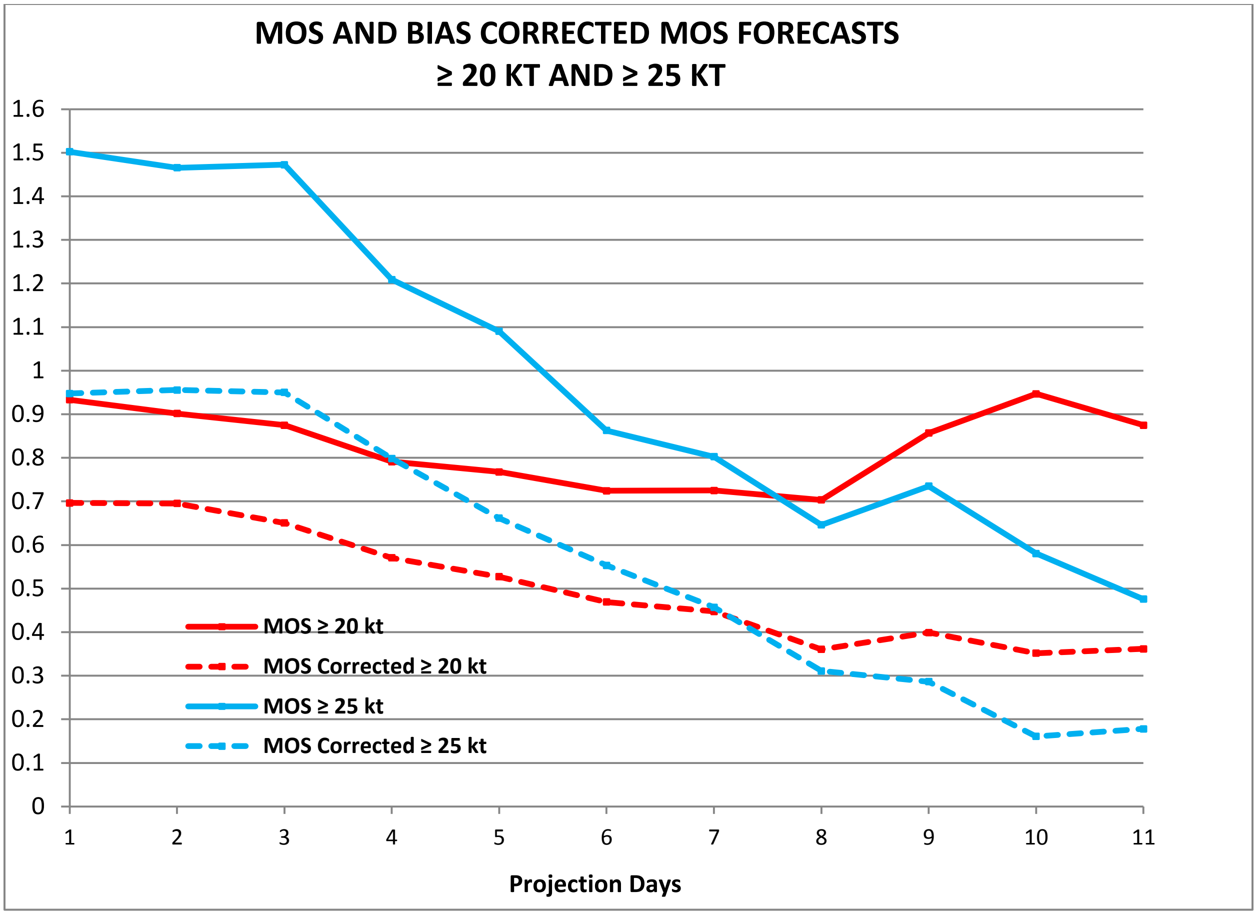

Figure 2. Bias of operational MOS forecasts ≥ 20 kt and ≥ 25 kt for the 6-month test period. Bias is defined as the number forecast in the category divided by the number observed in that category. Projections are shown at 24-h intervals (labeled in days for clarity).

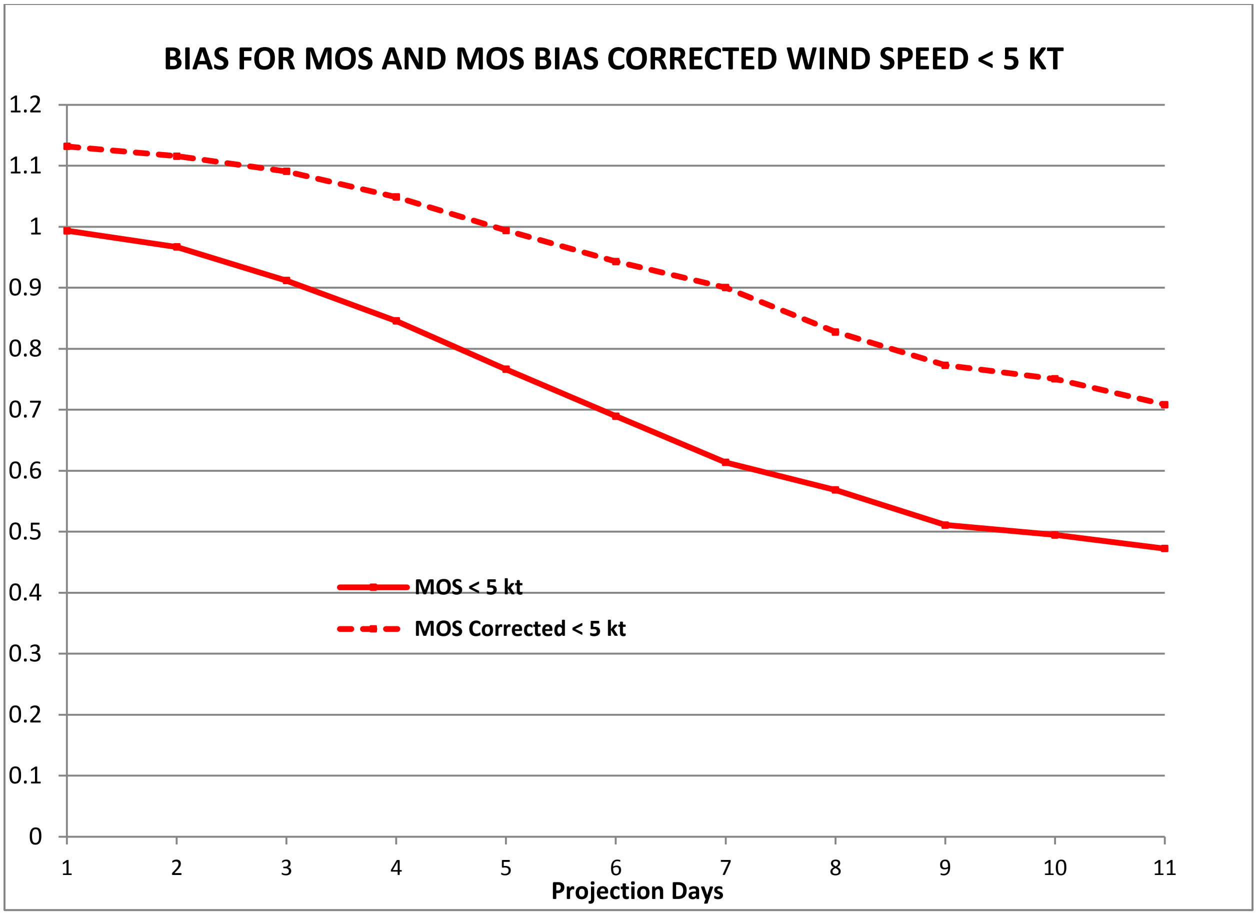

Figure 3. Same as Fig. 2 except for bias of forecasts < 5 kt.

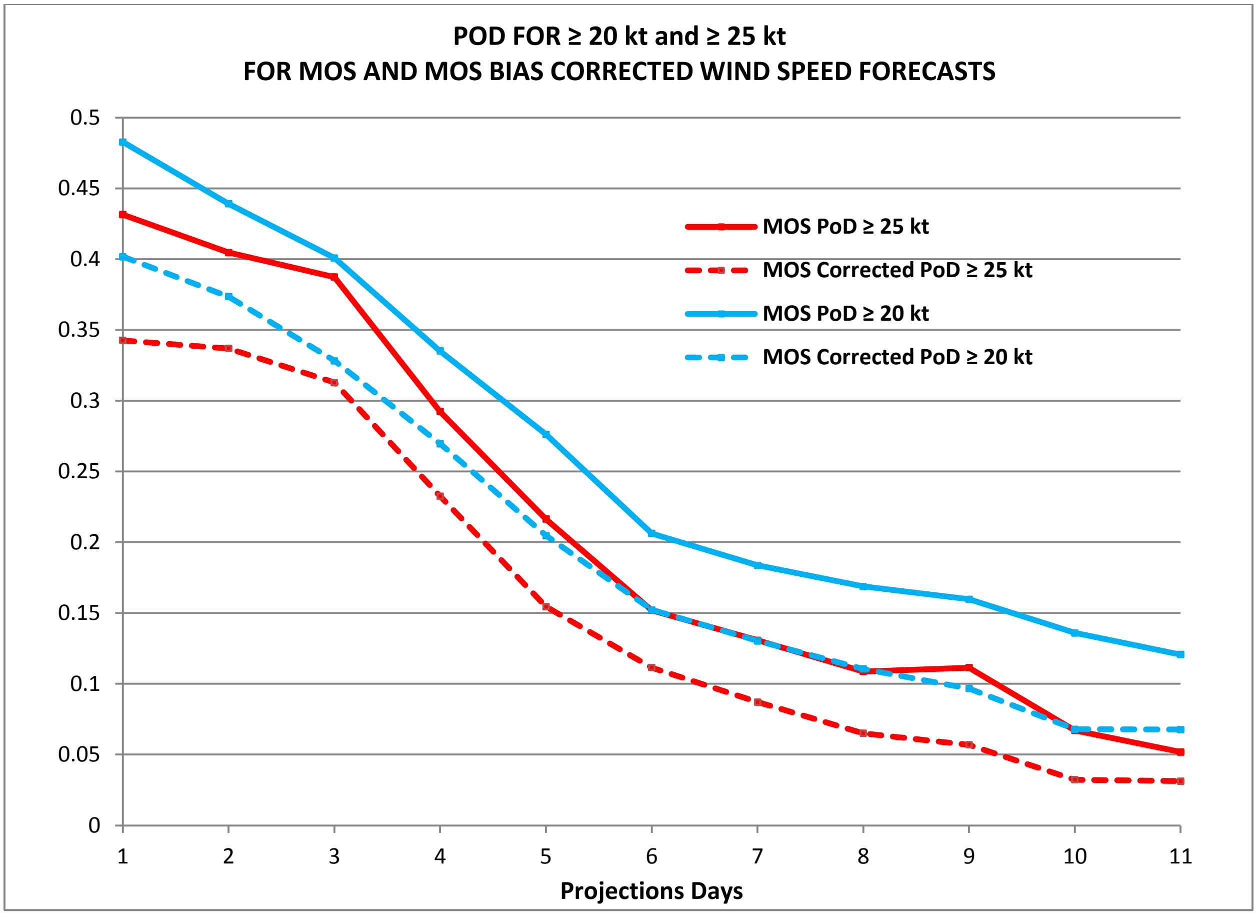

Figure 4. Same as Fig. 2 except for PoD for forecasts $\geq$ 20 kt and $\geq$ 25 kt.

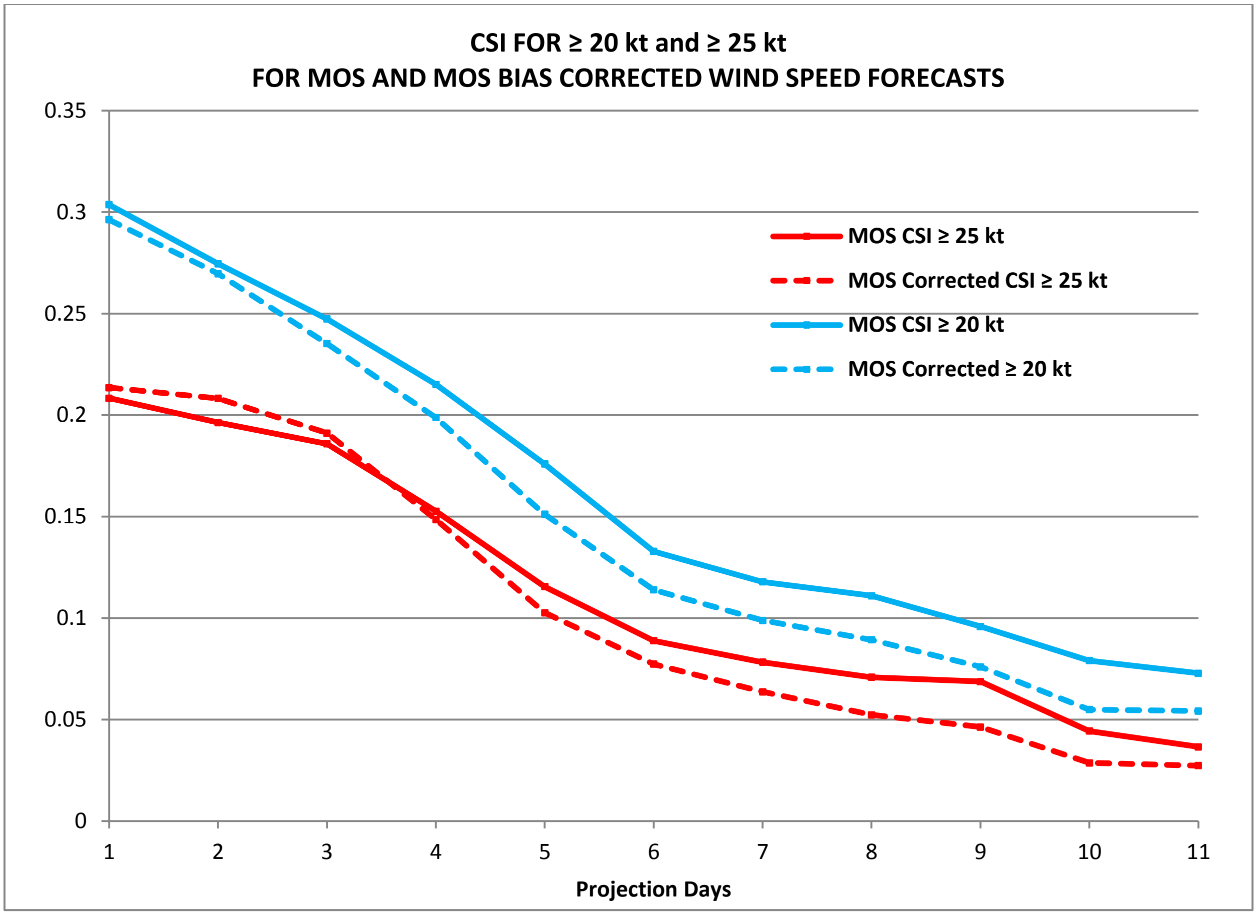

Figure 5. Same as Fig. 2 except for CSI for forecasts ≥ 20 kt and ≥ 25 kt.

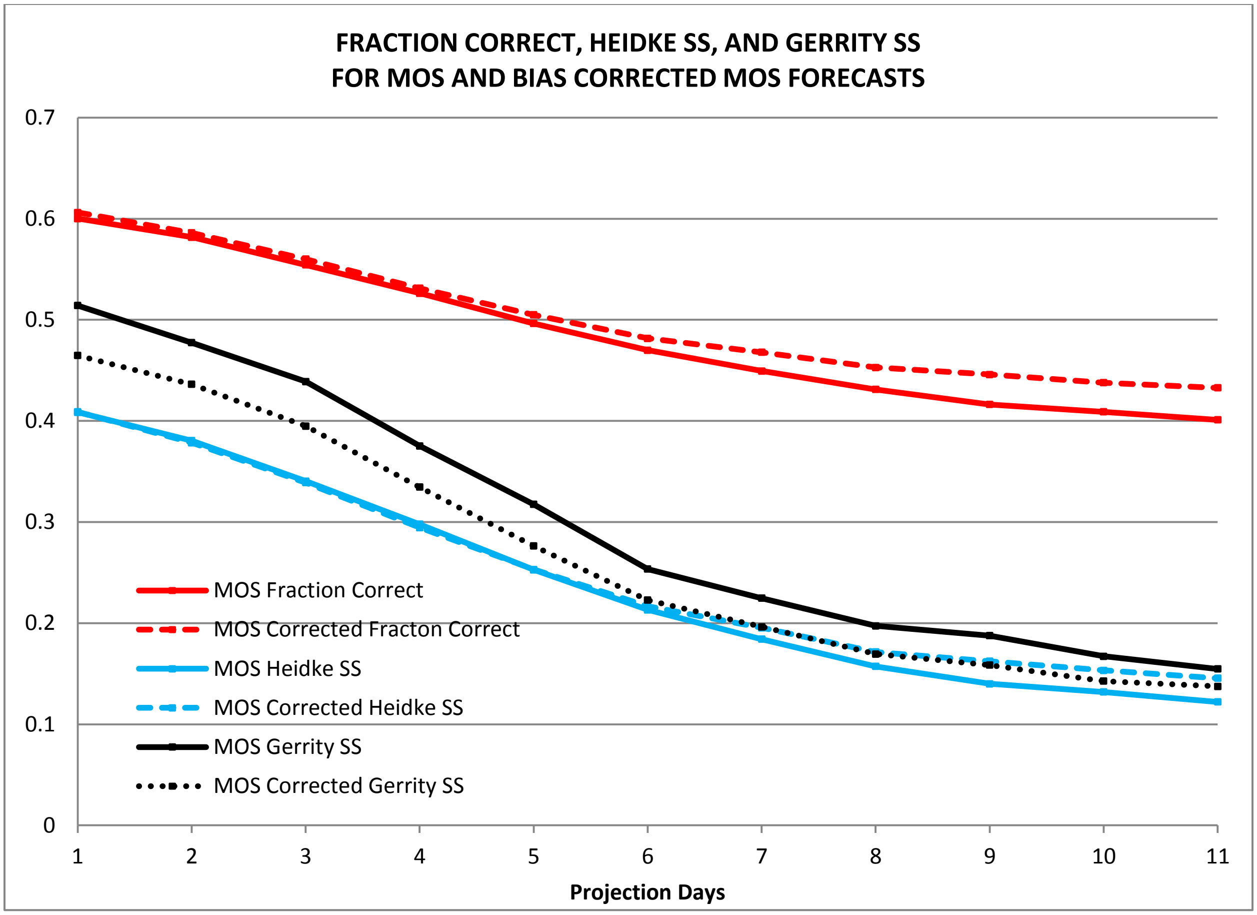

Figure 6. Same as Fig. 2 except for fraction correct, Heidke SS and Gerrity SS.

Model	GEFS	BC GEFS	MOS	MOS(I)	BC MOS(I)
MAE	3.47	3.04	2.56	2.69	2.64

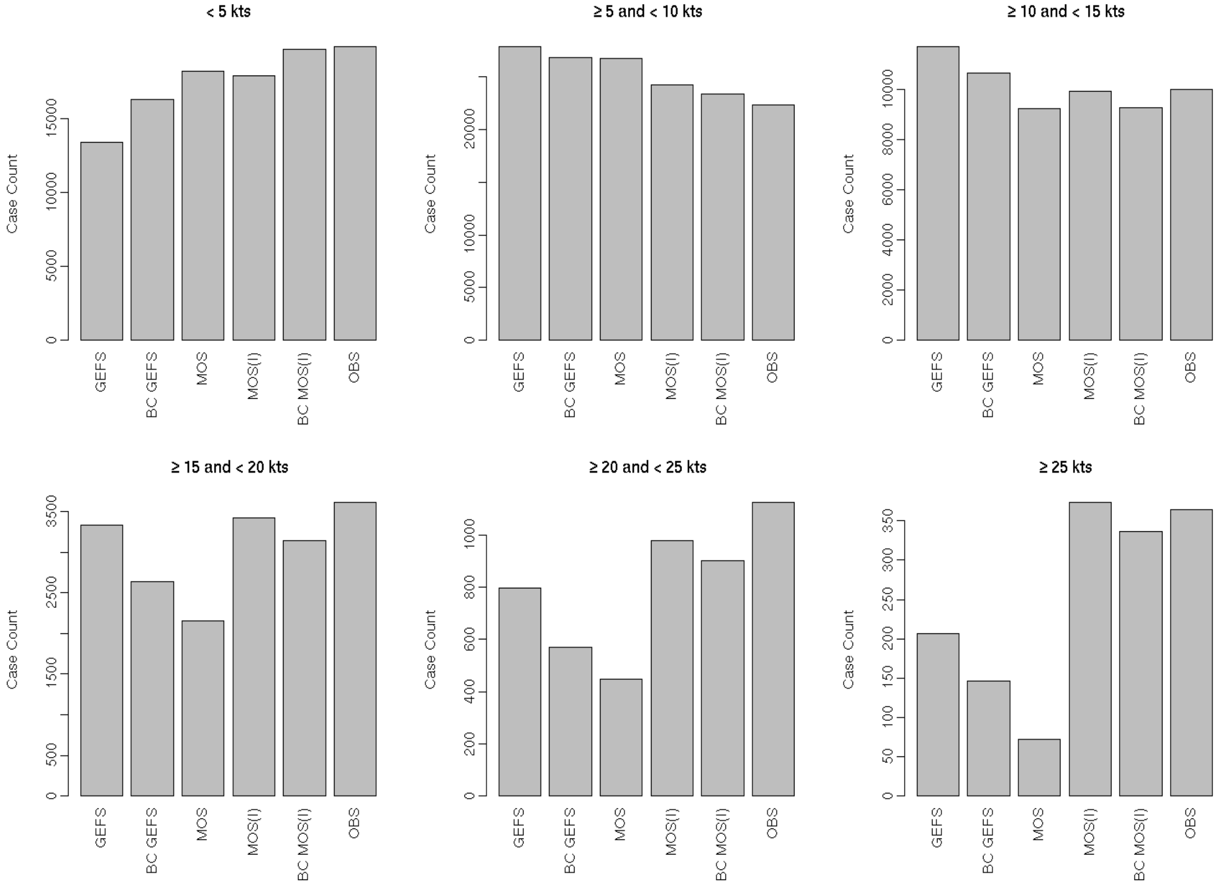

Figure 7. Cool season MAEs and frequencies for five different sets of 10-m, 48-h forecasts: (1) GEFS, (2) BC GEFS, (3) MOS based on GEFS, (4) partially inflated MOS forecasts, and (5) BC MOS partially inflated forecasts. Also shown in the bar graphs (rightmost column) are the corresponding frequencies for observations. Note that the ordinates are different for the different categories.

Model	GEFS	BC GEFS	MOS	MOS(I)	BC MOS(I)
MAE	3.76	3.50	3.10	3.31	3.24

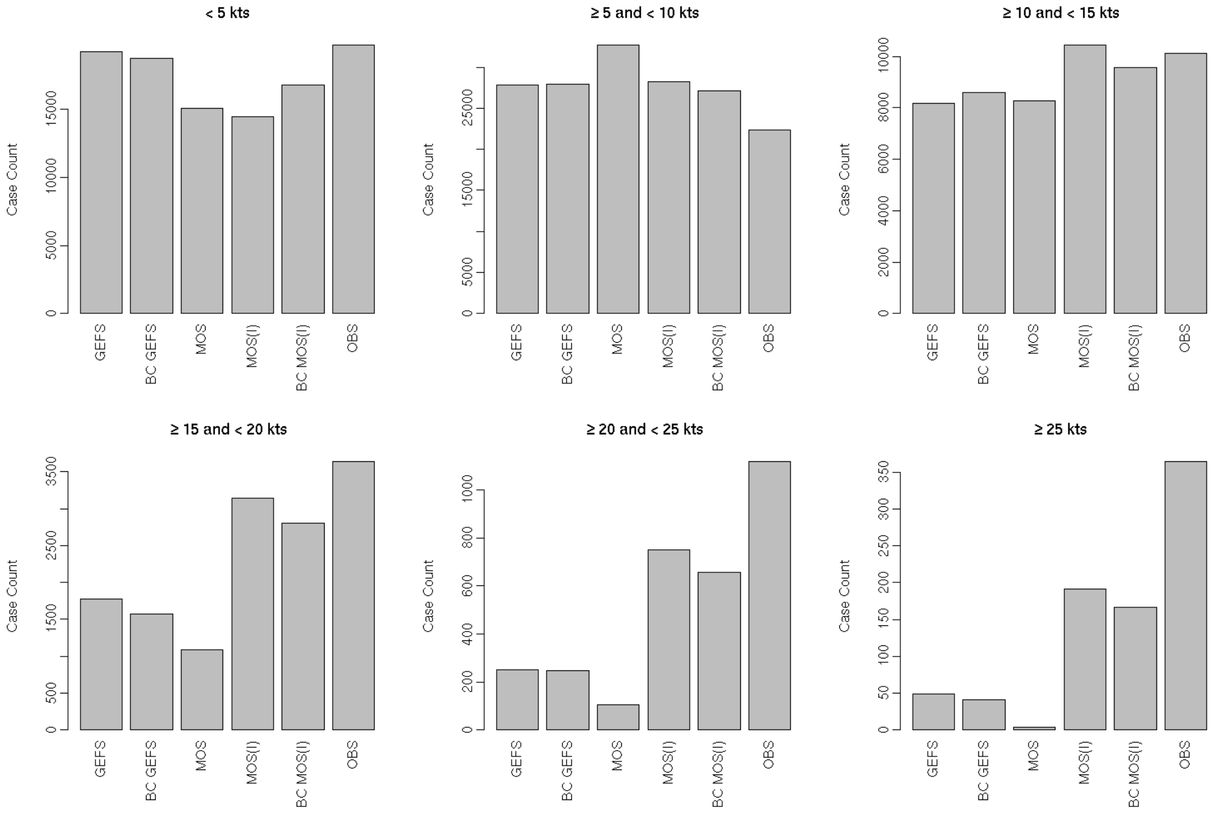

Figure 8. Same as Fig. 7 except for the 120-h projection.

Model	GEFS	BC GEFS	MOS	MOS(I)	BC MOS(I)
MAE	4.18	3.93	3.50	3.97	3.83

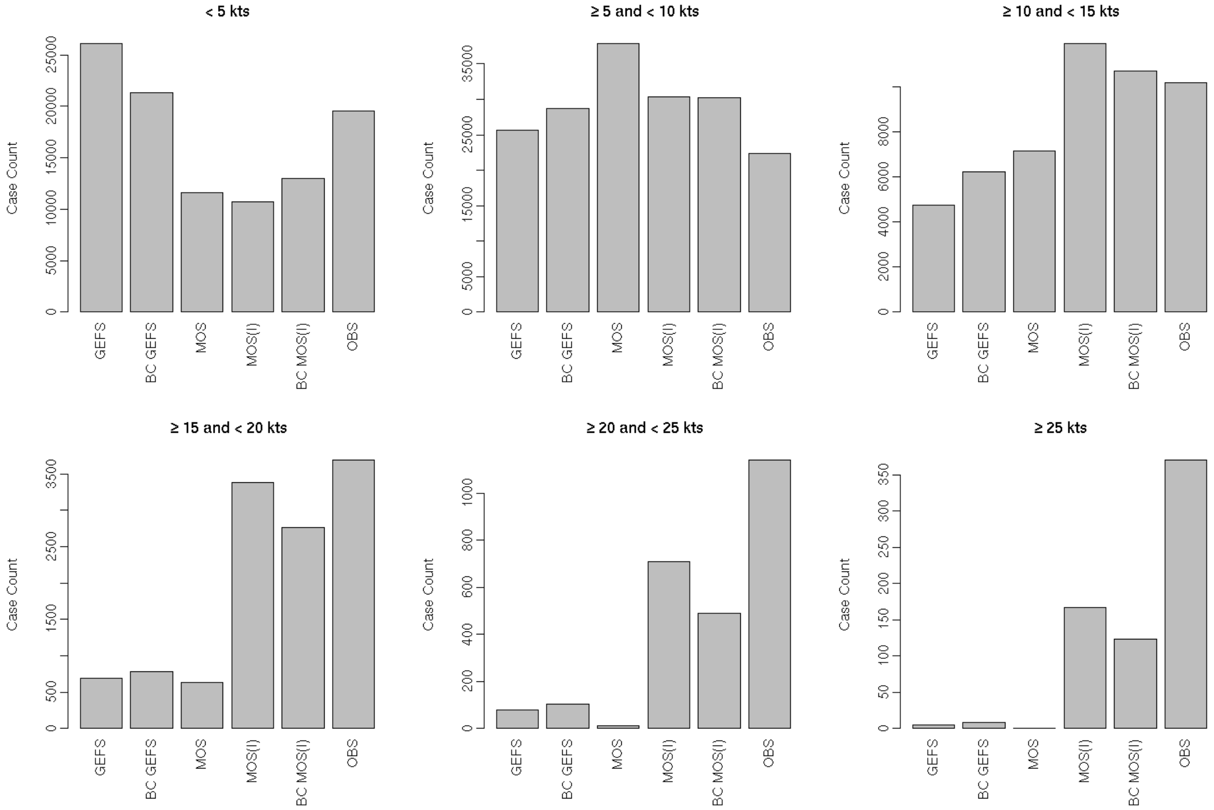

Figure 9. Same as Fig. 7 except for the 192-h projection.

Model	GEFS	BC GEFS	MOS	MOS(I)	BC MOS(I)
MAE	3.08	2.81	2.52	2.70	2.64

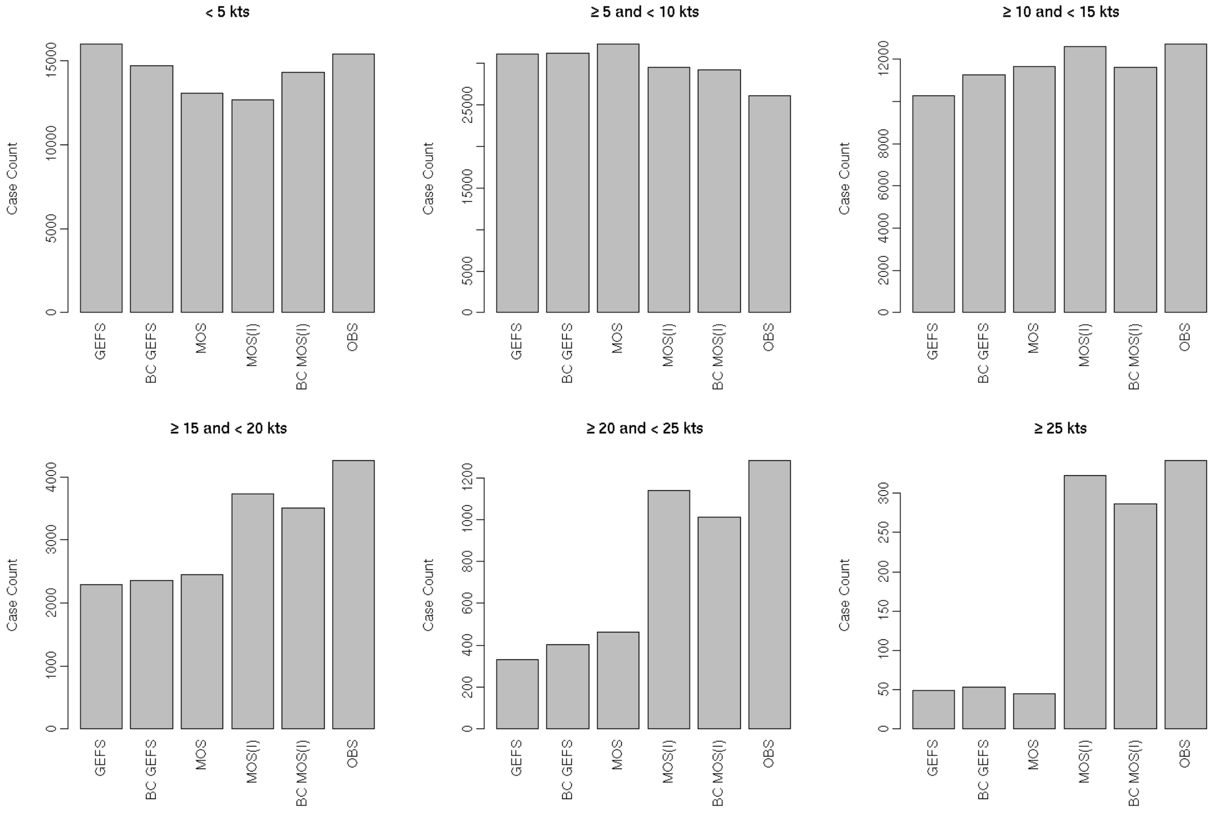

Figure 10. Same as Fig. 7 except for the 2013 warm season.

Model	GEFS	BC GEFS	MOS	MOS(I)	BC MOS(I)
MAE	3.44	3.13	2.80	3.03	2.96

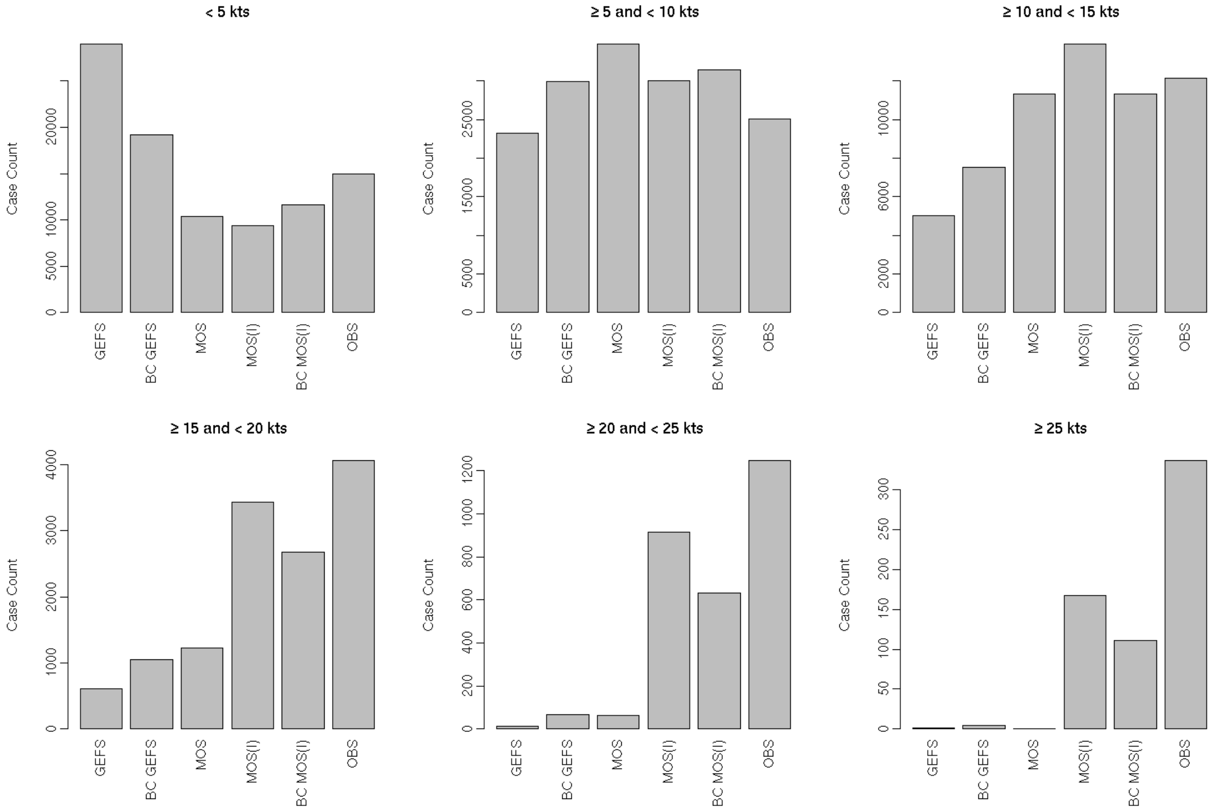

Figure 11. Same as Fig. 8 except for the 2013 warm season.

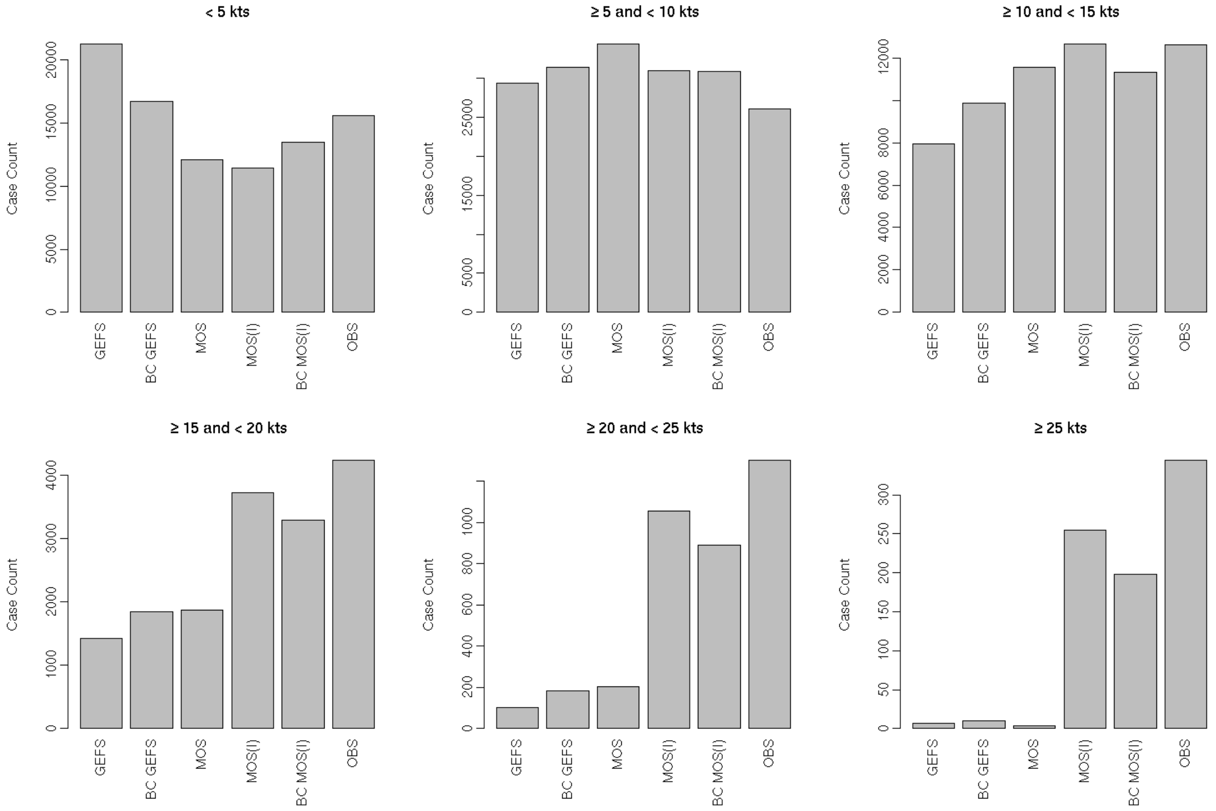

Figure 12. Same as Fig. 9 except for the 2013 warm season.